

(19)



JAPANESE PATENT OFFICE

PATENT ABSTRACTS OF JAPAN

(11) Publication number: **09319391 A**(43) Date of publication of application: **12.12.97**

(51) Int. Cl.

G10L 3/00
G10L 5/04
(21) Application number: **08250150**(22) Date of filing: **20.09.96**
(30) Priority: **12.03.96 JP 08 54714**
29.03.96 JP 08 77393
(71) Applicant: **TOSHIBA CORP**
(72) Inventor: **KAGOSHIMA TAKEHIKO**
AKAMINE MASAMI
(54) **SPEECH SYNTHESIZING METHOD**

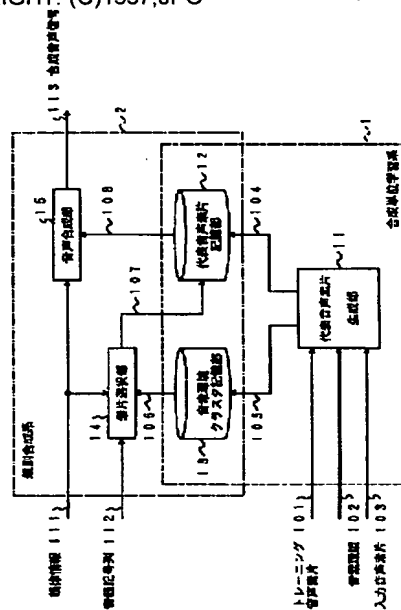
(57) Abstract:

PROBLEM TO BE SOLVED: To make it possible to effectively improve the tone quality of a synthesized speech by a text speech synthesis by forming such a representative phoneme that diminishes the distortion to natural speech at the level of the synthesized speech by taking the influence of change in pitch and the continuation time duration into consideration and synthesizing the speech by using the representative phoneme.

SOLUTION: The plural representative phonemes are formed by changing the pitch period and continuation time duration in a representative phoneme forming section 11. The distortion to the natural speeches is evaluated at the level of the synthesized speeches formed by changing at least either of the pitch and continuation time duration with respect to input phonemes 103. The phonemes selected from the input phonemes 103 based thereon are determined as the representative phonemes 104. The speech synthesis is executed in a speech synthesizing section 15 by connecting these representative phonemes 104, by which the high-quality synthesized speech signals 113

approximate to the natural speeches are formed.

COPYRIGHT: (C)1997,JPO



(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平9-319391

(43) 公開日 平成 9 年 (1997) 12 月 12 日

(51) Int. Cl. ⁶

G10L 3/00

5/04

識別記号

庁内整理番号

F I

G10L 3/00

5/04

技術表示箇所

H

F

審査請求 未請求 請求項の数12 O L (全25頁)

(21) 出願番号 特願平8-250150

(22) 出願日 平成 8 年 (1996) 9 月 20 日

(31) 優先権主張番号 特願平8-54714

(32) 優先日 平 8 (1996) 3 月 12 日

(33) 優先権主張国 日本 (J P)

(31) 優先権主張番号 特願平8-77393

(32) 優先日 平 8 (1996) 3 月 29 日

(33) 優先権主張国 日本 (J P)

(71) 出願人 000003078

株式会社東芝

神奈川県川崎市幸区堀川町72番地

(72) 発明者 籠嶋 岳彦

神奈川県川崎市幸区小向東芝町 1 番地 株

式会社東芝研究開発センター内

(72) 発明者 赤嶺 政巳

神奈川県川崎市幸区小向東芝町 1 番地 株

式会社東芝研究開発センター内

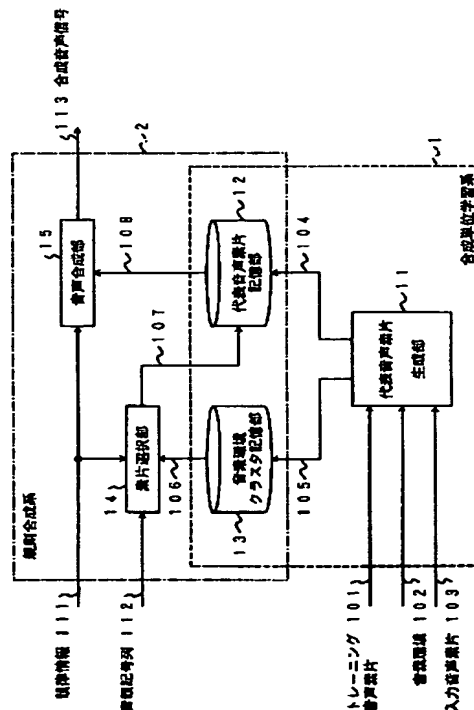
(74) 代理人 弁理士 鈴江 武彦 (外 6 名)

(54) 【発明の名称】 音声合成方法

(57) 【要約】

【課題】 テキスト音声合成による合成音声の音質を効果的に向上させることができる音声合成方法を提供する。

【解決手段】 代表音声素片生成部 11 において音素環境 102 がラベル付けされたトレーニング音声素片 101 のピッチ・継続時間長に従って入力音声素片 103 のピッチ・継続時間長を変更して複数の合成音声素片を生成し、合成音声素片とトレーニング音声素片 101 との間の距離尺度に基づいて入力音声素片 103 から代表音声素片 104 を選択して代表音声素片記憶部 12 に記憶し、さらに距離尺度に基づいて代表音声素片にそれぞれ対応する複数の音素環境クラスタ 105 を生成して音素環境クラスタ記憶部 13 に記憶し、代表音声素片記憶部 12 から入力音素の音素環境を含む音素環境クラスタに対応する代表音声素片を読み出して音声合成部 15 で接続することにより、合成音声信号 113 を生成する。



【特許請求の範囲】

【請求項 1】複数の第 1 の音声素片のピッチおよび継続時間長の少なくとも一方に従って複数の第 2 の音声素片のピッチおよび継続時間長の少なくとも一方を変更することにより複数の合成音声素片を生成し、これらの合成音声素片と前記第 1 の音声素片との間の距離尺度に基づいて前記第 2 の音声素片から複数の代表音声素片を選択して記憶し、これらの代表音声素片から所定の代表音声素片を選択して接続することによって音声を合成することを特徴とする音声合成方法。

【請求項 2】音素環境がラベル付けされた複数の第 1 の音声素片のピッチおよび継続時間長の少なくとも一方に従って複数の第 2 の音声素片のピッチおよび継続時間長の少なくとも一方を変更して複数の合成音声素片を生成し、これらの合成音声素片と前記第 1 の音声素片との間の距離尺度に基づいて前記第 2 の音声素片から複数の代表音声素片を選択して記憶し、前記距離尺度に基づいて前記代表音声素片にそれぞれ対応する複数の音素環境クラスタを生成し、前記代表音声素片から入力音素の音素環境を含む音素環境クラスタに対応する代表音声素片を選択して接続することによって音声を合成することを特徴とする音声合成方法。

【請求項 3】音素環境がラベル付けされた複数の第 1 の音声素片のピッチおよび継続時間長の少なくとも一方に従って複数の第 2 の音声素片のピッチおよび継続時間長の少なくとも一方を変更して複数の合成音声素片を生成し、これらの合成音声素片と前記第 1 の音声素片との間の距離尺度に基づいて複数の音素環境クラスタを生成し、前記距離尺度に基づいて前記第 2 の音声素片から各音素環境クラスタにそれぞれ対応する複数の代表音声素片を選択して記憶し、これらの代表音声素片から所定の代表音声素片を選択して接続することによって音声を合成することを特徴とする音声合成方法。

【請求項 4】音素環境がラベル付けされた複数の第 1 の音声素片のピッチおよび継続時間長の少なくとも一方に従って複数の第 2 の音声素片のピッチおよび継続時間長の少なくとも一方を変更して複数の合成音声素片を生成し、これらの合成音声素片と前記第 1 の音声素片との間の距離尺度に基づいて複数の音素環境クラスタを生成し、前記距離尺度に基づいて前記第 2 の音声素片から各音素環境クラスタにそれぞれ対応する複数の代表音声素片を選択して記憶し、これらの代表音声素片から入力音素の音素環境を含む音素環境クラスタに対応する代表音声素片を選択して接続

することによって音声を合成することを特徴とする音声合成方法。

【請求項 5】複数の第 1 の音声素片のピッチおよび継続時間長の少なくとも一方に従って複数の第 2 の音声素片のピッチおよび継続時間長の少なくとも一方を変更することにより複数の合成音声素片を生成し、これらの合成音声素片についてスペクトル整形を行い、このスペクトル整形を行った後の各合成音声素片と前記第 1 の音声素片との間の距離尺度に基づいて前記第 2 の音声素片から複数の代表音声素片を選択して記憶し、これらの代表音声素片から所定の代表音声素片を選択して接続することによって音声を合成し、この合成した音声のスペクトル整形を行って最終的な合成音声を生成することを特徴とする音声合成方法。

【請求項 6】音素環境がラベル付けされた複数の第 1 の音声素片のピッチおよび継続時間長の少なくとも一方に従って複数の第 2 の音声素片のピッチおよび継続時間長の少なくとも一方を変更して複数の合成音声素片を生成し、これらの合成音声素片についてスペクトル整形を行い、このスペクトル整形を行った後の各合成音声素片と前記第 1 の音声素片との間の距離尺度に基づいて前記第 2 の音声素片から複数の代表音声素片を選択して記憶し、前記距離尺度に基づいて前記代表音声素片にそれぞれ対応する複数の音素環境クラスタを生成し、前記代表音声素片から入力音素の音素環境を含む音素環境クラスタに対応する代表音声素片を選択して接続することによって音声を合成し、この合成した音声のスペクトル整形を行って最終的な合成音声を生成することを特徴とする音声合成方法。

【請求項 7】音素環境がラベル付けされた複数の第 1 の音声素片のピッチおよび継続時間長の少なくとも一方に従って複数の第 2 の音声素片のピッチおよび継続時間長の少なくとも一方を変更して複数の合成音声素片を生成し、これらの合成音声素片についてスペクトル整形を行い、このスペクトル整形を行った後の各合成音声素片と前記第 1 の音声素片との間の距離尺度に基づいて複数の音素環境クラスタを生成し、前記距離尺度に基づいて前記第 2 の音声素片から各音素環境クラスタにそれぞれ対応する複数の代表音声素片を選択して記憶し、これらの代表音声素片から所定の代表音声素片を選択して接続することによって音声を合成し、この合成した音声のスペクトル整形を行って最終的な合成音声を生成することを特徴とする音声合成方法。

【請求項 8】音素環境がラベル付けされた複数の第 1 の音声素片のピッチおよび継続時間長の少なくとも一方に従って複数の第 2 の音声素片のピッチおよび継続時間長の少なくとも一方を変更して複数の合成音声素片を生成

し、
これらの合成音声素片についてスペクトル整形を行い、
このスペクトル整形を行った後の各合成音声素片と前記
第 1 の音声素片との間の距離尺度に基づいて複数の音素
環境クラスタを生成し、
前記距離尺度に基づいて前記第 2 の音声素片から各音素
環境クラスタにそれぞれ対応する複数の代表音声素片を
選択して記憶し、
これらの代表音声素片から入力音素の音素環境を含む音
素環境クラスタに対応する代表音声素片を選択して接続
することによって音声合成し、
この合成した音声のスペクトル整形を行って最終的な合
成音声の生成することを特徴とする音声合成方法。

【請求項 9】前記代表音声素片として、音源信号と該音
源信号を入力として合成音声信号を生成する合成フィル
タの係数の組の情報を記憶することを特徴とする請求項
1 ～ 8 のいずれか 1 項に記載の音声合成方法。

【請求項 10】前記音源信号と前記合成フィルタの係数
を量子化し、これら量子化した音源信号と合成フィルタ
の係数の組の情報を前記代表音声素片として記憶するこ
とを特徴とする請求項 9 に記載の音声合成方法。

【請求項 11】前記代表音声素片として、音源信号と該
音源信号を入力として合成音声信号を生成する合成フィ
ルタの係数の組の情報を記憶し、
かつ該代表音声素片の情報として記憶する音源信号およ
び合成フィルタの係数のうちの少なくとも一方の数が音
声合成単位の総数より少ないことを特徴とする請求項 1
～ 8 のいずれか 1 項に記載の音声合成方法。

【請求項 12】前記代表音声素片として、音源信号と該
音源信号を入力として合成音声信号を生成する合成フィ
ルタの係数の組の情報を記憶し、
かつ該代表音声素片の情報として記憶する音源信号およ
び合成フィルタの係数のうちの少なくとも一方の数が前
記音素環境クラスタの総数より少ないことを特徴とする
請求項 2、3、6、7、8、9 のいずれか 1 項に記載の
音声合成方法。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、テキスト音声合成
のための音声合成方法に係り、特に音韻記号列、ピッチ
および音韻継続時間長などの情報から音声信号を生成す
る音声合成方法に関する。

【0002】

【従来の技術】任意の文章から人工的に音声信号を作り
出すことをテキスト音声合成という。テキスト音声合成
は、一般的に言語処理部、音韻処理部および音声合成部
の 3 つの段階によって行われる。入力されたテキスト
は、まず言語処理部において形態素解析や構文解析など
が行われ、次に音韻処理部においてアクセントやイント
ネーションの処理が行われて、音韻記号列・ピッチ・音

韻継続時間長などの情報が出力される。最後に、音声信
号合成部で音韻記号列・ピッチ・音韻継続時間長などの
情報から音声信号を合成する。そこで、テキスト音声合
成に用いる音声合成方法は、任意の音韻記号列を任意の
韻律で音声合成することが可能な方法でなければならない。

【0003】このような任意の音韻記号列を音声合成す
る音声合成装置の原理は、母音を V、子音を C で表す
と、CV、CVC、VCV といった基本となる小さな単
位の特徴パラメータ（これを代表音声素片という）を記
憶し、これらを選択的に読み出した後、ピッチや継続時
間長を制御して接続することにより、音声合成すると
いうものである。従って、記憶されている代表音声素片
が合成音声の品質を大きく左右することになる。

【0004】従来、これらの代表音声素片の作成はもっ
ぱら人手に頼っており、音声信号の中から試行錯誤的に
切り出してくる場合がほとんどであるため、膨大な労力
を要していた。このような代表音声素片作成の作業を自
動化し、音声合成に使用するのに適した代表音声素片を
容易に生成する方法として、例えば音素環境クラスタリ
ング（COC）と呼ばれる技術が特開昭 64 - 78300
「音声合成方法」に開示されている。

【0005】COC の原理は、音素名や音素環境のラベ
ルを多数の音声素片に付与し、そのラベルが付与された
音声素片を音声素片間の距離尺度に基づいて音素環境に
関する複数のクラスタに分類し、その各クラスタのセン
トロイドを代表音声素片とするものである。ここで、音
素環境とは当該音声素片にとっての環境となる要因全て
の組合せであり、その要因としては当該音声素片の音素
名、先行音素、後続音素、後々続音素、ピッチ周期、パ
ワー、ストレスの有無、アクセント核からの位置、息継
ぎからの時間、発声速度、感情などが考えられる。実音
声中の各音素は音素環境によって音韻が変化しているた
め、音素環境に関する複数のクラスタ毎に代表音声素片
を記憶しておくことにより、音素環境の影響を考慮した
自然な音声合成することが可能となっている。

【0006】

【発明が解決しようとする課題】上に述べたように、テ
キスト音声合成のための音声合成では、代表音声素片の
ピッチや継続時間長を指定された値に変更して合成する
必要がある。このようなピッチや継続時間長の変更によ
り、代表音声素片を切り出してきた音声信号の音質と比
較して合成音声の音質がある程度劣化することになる。

【0007】これに対して、上記の COC によるクラス
タリングでは、音声素片間の距離尺度に基づいてクラス
タリングを行っているにすぎないため、合成の際のピッ
チや継続時間の変更の効果が全く考慮されていないとい
う問題がある。すなわち、COC によるクラスタリング
および各クラスタの代表音声素片は、実際にピッチや継
続時間長を変更して合成された合成音声のレベルでは、

必ずしも適当なものになっているという保証はない。

【0008】本発明は、このような問題点を解決すべくなされたものであり、テキスト音声合成による合成音声の音質を効果的に向上させることができる音声合成方法を提供することを目的とする。

【0009】

【課題を解決するための手段】上記の課題を解決するため、本発明はピッチや継続時間長の変更の影響を考慮して、合成音声のレベルで自然音声に対する歪みが小さくなるような代表音声素片を生成し、その代表音声素片を用いて音声を合成することにより、自然音声に近い合成音声を生成するようにしたものである。

【0010】すなわち、本発明に係る音声合成方法は、複数の第1の音声素片のピッチおよび継続時間長の少なくとも一方に従って複数の第2の音声素片のピッチおよび継続時間長の少なくとも一方を変更することにより複数の合成音声素片を生成し、これらの合成音声素片と第1の音声素片との間の距離尺度に基づいて第2の音声素片から複数の代表音声素片を選択して記憶し、これらの代表音声素片から所定の代表音声素片を選択して接続することによって音声を合成することを特徴とする。

【0011】ここで、第1および第2の音声素片は、CV、VCV、CVCといった音声合成単位で音声信号中から切り出される素片であり、切り出された波形もしくはその波形から何らかの方法で抽出されたパラメータ系列などを表すものとする。これらのうち、第1の音声素片は合成音声の歪みを評価するために用いられ、また第2の音声素片は代表音声素片の候補として用いられる。合成音声素片は、第2の音声素片に対して少なくともピッチまたは継続時間長を変更して生成される合成音声波形またはパラメータ系列などを表す。

【0012】合成音声素片と第1の音声素片との間の距離尺度によって、合成音声の歪みが表わされる。従って、この距離尺度つまり歪みがより小さくなる音声素片を第2の音声素片から選択して代表音声素片として記憶しておき、これらの代表音声素片から所定の代表音声素片を選択して接続すれば、自然音声に近い高品質の合成音声が生成される。

【0013】本発明の第1の態様では、音素環境がラベル付けされた複数の第1の音声素片のピッチおよび継続時間長の少なくとも一方に従って複数の第2の音声素片のピッチおよび継続時間長の少なくとも一方を変更して複数の合成音声素片を生成し、これらの合成音声素片と第1の音声素片との間の距離尺度に基づいて第2の音声素片から複数の代表音声素片を選択して記憶し、前記距離尺度に基づいて代表音声素片にそれぞれ対応する複数の音素環境クラスタを生成し、複数の代表音声素片から入力音素の音素環境を含む音素環境クラスタに対応する代表音声素片を選択して接続することによって音声を合成する。

【0014】ここで、音素環境とは前述した通り音声素片にとっての環境となる要因、例えば当該音声素片の音素名、先行音素、後続音素、後々続音素、ピッチ周期、パワー、ストレスの有無、アクセント核からの位置、息継ぎからの時間、発声速度、感情といった要素の組み合わせであり、音素環境クラスタとは言い換えれば音素環境の集合であり、例えば「当該素片の音韻が／k a／、先行音韻が／i／または／u／、ピッチ周波数が200 Hz以下」というようなものを意味する。

【0015】第1の態様のように、距離尺度つまり合成音声の歪みに基づいて代表音声素片にそれぞれ対応する複数の音素環境クラスタを生成し、入力音素の音素環境を含む音素環境クラスタに対応する代表音声素片を選択して接続するようにすれば、例えば同一音素名の音声素片が複数の音素環境に存在する場合でも、実際の入力音素の音素環境が含まれる音素環境クラスタに対応する代表音声素片のみが選択されることにより、より自然な合成音声が得られる。

【0016】本発明の第2の態様では、音素環境がラベル付けされた複数の第1の音声素片のピッチおよび継続時間長の少なくとも一方に従って複数の第2の音声素片のピッチおよび継続時間長の少なくとも一方を変更して複数の合成音声素片を生成し、これらの合成音声素片と第1の音声素片との間の距離尺度に基づいて複数の音素環境クラスタを生成し、前記距離尺度に基づいて第2の音声素片から各音素環境クラスタにそれぞれ対応する複数の代表音声素片を選択して記憶し、これらの代表音声素片から所定の代表音声素片を選択して接続することによって音声を合成する。この第2の態様は、音声素片が一つの音素環境にのみ存在する場合に有効である。

【0017】本発明の第3の態様では、音素環境がラベル付けされた複数の第1の音声素片のピッチおよび継続時間長の少なくとも一方に従って複数の第2の音声素片のピッチおよび継続時間長の少なくとも一方を変更して複数の合成音声素片を生成し、これらの合成音声素片と第1の音声素片との間の距離尺度に基づいて複数の音素環境クラスタを生成し、第1の音声素片と合成音声素片との間の距離尺度に基づいて第2の音声素片から各音素環境クラスタにそれぞれ対応する複数の代表音声素片を選択して記憶し、これらの代表音声素片から入力音素の音素環境を含む音素環境クラスタに対応する代表音声素片を選択して接続することによって音声を合成する。

【0018】この第3の態様によっても、第1の態様と同様に、例えば同一音素名の音声素片が複数の音素環境に存在する場合、実際の入力音素の音素環境が含まれる音素環境クラスタに対応する代表音声素片のみが選択されることにより、より自然な合成音声が得られる。

【0019】また、本発明に係る他の音声合成方法は、複数の第1の音声素片のピッチおよび継続時間長の少なくとも一方に従って複数の第2の音声素片のピッチおよ

び継続時間長の少なくとも一方を変更することにより複数の合成音声素片を生成し、さらにこれらの合成音声素片についてスペクトル整形を行い、このスペクトル整形を行った後の各合成音声素片と第 1 の音声素片との間の距離尺度に基づいて第 2 の音声素片から複数の代表音声素片を選択して記憶し、これらの代表音声素片から所定の代表音声素片を選択して接続することによって音声を合成し、この合成した音声のスペクトル整形を行って最終的な合成音声を生成することの特徴とする。

【0020】この場合、先に示した第 1、第 2 および第 3 の態様においても、複数の合成音声素片を生成した後、スペクトル整形を行うようにする。ここで、スペクトル整形は「めりはり」のある明瞭な音声を合成するための処理であり、例えばホルマント強調やピッチ強調を行う適応ポストフィルタによるフィルタリングによって実現される。

【0021】このように代表音声素片の接続によって合成される音声に対してスペクトル整形を行うと共に、合成音声素片に対しても同様のスペクトル整形を行うことによって、スペクトル整形後の最終的な合成音声のレベルで、自然音声に対する歪が小さくなるような代表音声素片を生成できるため、「めりはり」に優れたより明瞭な合成音声を得られる。

【0022】本発明においては、代表音声素片として、音源信号と該音源信号を入力として合成音声信号を生成する合成フィルタの係数の組の情報を記憶するようにしてもよい。この場合、音源信号と合成フィルタの係数を量子化し、これら量子化した音源信号と合成フィルタの係数の組の情報を記憶するようにすれば、代表音声素片として記憶する音源信号と合成フィルタの係数の数を減少させることができるため、合成単位の学習に要する計算時間が短縮され、かつ実際の音声合成時に必要なメモリ量が低減される。

【0023】さらに、代表音声素片の情報として記憶する音源信号および合成フィルタの係数のうちの少なくとも一方の数を音声合成単位の総数や、音素環境クラスタの総数より少なくすることも可能であり、このようにしても良好な合成音声を得ることができる。

【0024】

【発明の実施の形態】以下、図面を参照して本発明の実施形態を説明する。

(第 1 の実施形態) 図 1 は、本発明の第 1 の実施形態に係る音声合成方法を実現する音声合成装置の構成を示すブロック図である。この音声合成装置は、大きく分けて合成単位学習系 1 と規則合成系 2 からなる。実際にテキスト音声合成を行う場合に動作するのは規則合成系 2 であり、合成単位学習系 1 は事前に学習を行って代表音声素片を生成するものである。

【0025】まず、合成単位学習系 1 について説明する。合成単位学習系 1 は、代表音声素片とこれに付随す

る音素環境クラスタを生成する代表音声素片生成部 11 と代表音声素片記憶部 12 および音素環境クラスタ記憶部 13 により構成される。代表音声素片生成部 11 には、第 1 の音声素片であるトレーニング音声素片 101 とこれにラベル付けされた音素環境 102 および第 2 の音声素片である入力音声素片 103 が入力される。

【0026】代表音声素片生成部 11 では、トレーニング音声素片 101 にラベル付けされた音素環境 102 に含まれるピッチ周期および継続時間長の情報に従って、入力音声素片 103 のピッチ周期および継続時間長を変更することで複数の合成音声素片が内部的に生成され、さらにこれらの合成音声素片とトレーニング音声素片 101 との距離尺度に従って、代表音声素片 104 と音素環境クラスタ 105 が生成される。音素環境クラスタ 105 は、トレーニング音声素片 101 を後述するように音素環境に関するクラスタに分類して生成される。

【0027】代表音声素片 104 は代表音声素片記憶部 12 に記憶され、音素環境クラスタ 105 は代表音声素片 104 と対応付けられて音素環境クラスタ記憶部 13 に記憶される。代表音声素片生成部 11 の処理については、後に詳細に説明する。

【0028】次に、規則合成系 2 について説明する。規則合成系 2 は、代表音声素片記憶部 12 と音素環境クラスタ記憶部 13 と素片選択部 14 および音声合成部 15 により構成され、代表音声素片記憶部 12 と音素環境クラスタ記憶部 13 を合成単位学習系 1 と共有している。

【0029】素片選択部 14 には、入力音素の情報として、例えばテキスト音声合成のために入力テキストの形態素解析・構文解析後さらにアクセントやイントネーション処理を行って得られた韻律情報 111 と音韻記号列 112 が入力される。韻律情報 111 には、ピッチパターンおよび音韻継続時間長が含まれている。素片選択部 14 では、これらの韻律情報 111 と音韻記号列 112 から入力音素の音素環境を内部的に生成する。

【0030】そして、素片選択部 14 は音素環境クラスタ記憶部 13 より読み出された音素環境クラスタ 106 を参照して、入力音素の音素環境がどの音素環境クラスタに属するかを探索し、探索した音素環境クラスタに対応する代表音声素片選択情報 107 を代表音声素片記憶部 12 へ出力する。

【0031】音声合成部 15 は、代表音声素片選択情報 107 に従って代表音声素片記憶部 12 より選択的に読み出された代表音声素片 108 に対して、韻律情報 111 に従ってピッチ周期および音韻継続時間長を変更するとともに、素片の接続を行って合成音声信号 113 を出力する。ここで、ピッチおよび継続時間長を変更して素片を接続し音声を合成する方法としては、例えば残差駆動 LSP 方法や波形編集方法など公知の技術を用いることができる。

【0032】次に、本発明の特徴をなす代表音声素片生

成部 11 の処理手順について具体的に説明する。図 2 のフローチャートは、代表音声素片生成部 11 の第 1 の処理手順を示している。

【0033】この第 1 の実施形態による代表音声素片生成処理では、まず準備段階として連続発声された多数の音声データに対して音韻毎にラベリングを行い、CV、VCV、CVC などの合成単位に従って、トレーニング音声素片 T_i ($i = 1, 2, 3, \dots, N_t$) を切り出す。また、各トレーニング音声素片 T_i に対応する音素環境 P_i ($i = 1, 2, 3, \dots, N_t$) も抽出しておく。ただし、 N_t はトレーニング音声素片の個数を表す。音素環境 P_i は、少なくともトレーニング音声素片 T_i の音韻とそのピッチおよび継続時間長の情報を含むものとし、その他に必要に応じて前後の音素などの情報を含むものとする。

【0034】次に、上述したトレーニング音声素片 T_i の作成と同様の方法により、多数の入力音声素片 S_j ($j = 1, 2, 3, \dots, N_s$) を作成する。ただし、 N_s は入力音声素片の個数を表す。ここで、入力音声素片 S_j としてはトレーニング音声素片 T_i と同じものを使用してもよいし（すなわち $T_i = S_j$ ）、トレーニング音声素片 T_i とは異なる音声素片を作成してもよい。いずれにしても、豊富な音韻環境を有する多数のトレーニング音声素片および入力音声素片が用意されていることが望ましい。

【0035】このような準備段階を経た後、まず音声合成ステップ S21 で、音素環境 P_i に含まれるピッチおよび継続時間長に等しくなるように、入力音声素片 S_j のピッチおよび継続時間長を変更して音声合成することにより、合成音声素片 G_{ij} を生成する。ここでのピッチおよび継続時間長の変更は、音声合成部 15 におけるピッチおよび継続時間長の変更と同様の方法で行われる

$$E_{01}(U) = \sum_{i=1}^{N_t} \min(e_{i11}, e_{i12}, e_{i13}, \dots, e_{i1N}) \quad (1)$$

【0040】ただし、 $\min(e_{i11}, e_{i12}, e_{i13}, \dots, e_{i1N})$ は $e_{i11}, e_{i12}, e_{i13}, \dots, e_{i1N}$ の中の最小値を表す関数である。集合 U の組合せは $N_s! / \{N! (N_s - N)!\}$ 通りあり、これらの音声素片の集合 U の中から評価関数 $E_{01}(U)$ を最小にする U を探索し、その要素 u_k を代表音声素片 D_k とする。

【0041】最後に、音素環境クラスタ生成ステップ S

$$E_{c1} = \sum_{k=1}^N \sum_{P_i \in C_k} e_{ijk}$$

【0043】こうしてステップ S23 および S24 で生成された代表音声素片 D_k および音素環境クラスタ C_k は、図 1 の代表音声素片記憶部 12 および音素環境クラ

ものとする。全ての音素環境 P_i ($i = 1, 2, 3, \dots, N_t$) に従って入力音声素片 S_j ($j = 1, 2, 3, \dots, N_s$) を用いて音声の合成を行うことにより、 $N_t \times N_s$ 個の合成音声素片 G_{ij} ($i = 1, 2, 3, \dots, N_t, j = 1, 2, 3, \dots, N_s$) を生成する。

【0036】次に、歪み評価ステップ S22 では、合成音声素片 G_{ij} の歪み e_{ij} の評価を行う。この歪み e_{ij} の評価は、合成音声素片 G_{ij} とトレーニング音声素片 T_i との間の距離尺度を求めることにより行う。距離尺度には、何らかのスペクトル距離を用いることができる。例えば、合成音声素片 G_{ij} およびトレーニング音声素片 T_i について、FFT（高速フーリエ変換）などを用いてパワースペクトルを求めて各パワースペクトル間の距離を評価する方法や、あるいは線形予測分析を行って LP C または LSP パラメータなどを求めて各パラメータ間の距離を評価する方法などがある。その他にも、短時間フーリエ変換やウェーブレット変換などの変換係数を用いて評価する方法も用いることができる。また、各素片のパワーを正規化した上で歪みの評価を行う方法でもよい。

【0037】次に、代表音声素片生成ステップ S23 では、ステップ S22 で得られた歪み e_{ij} に基づいて、入力音声素片 S_j の中から指定された代表音声素片数 N の代表音声素片 D_k ($k = 1, 2, 3, \dots, N$) を選択する。

【0038】代表音声素片選択法の一例を説明する。入力音声素片 S_j の中から選択された N 個の音声素片の集合 $U = \{u_k \mid u_k = S_j \text{ (} k = 1, 2, 3, \dots, N \text{)}\}$ に対して、歪みの総和を表す評価関数 $E_{01}(U)$ を次式 (1) のように定義する。

【0039】

【数 1】

24 では、音素環境 P_i 、歪み e_{ij} および代表音声素片 D_k より、音素環境に関する複数のクラスタ（音素環境クラスタ） C_k ($k = 1, 2, 3, \dots, N$) を生成する。音素環境クラスタ C_k は、例えば次式 (2) で表されるクラスタリングの評価関数 E_{c1} を最小化するクラスタを探索することによって得られる。

【0042】

【数 2】

(2)

スタ記憶部 13 にそれぞれ記憶される。

【0044】次に、図 3 のフローチャートを参照して代表音声素片生成部 11 の第 2 の処理手順について説明す

る。この第2の処理手順による代表音声素片生成処理では、まず初期音素環境クラスタ生成ステップS30において、何らかの先見的な知識に基づいて予め音素環境のクラスタリングを行い、初期音素環境クラスタを生成する。音素環境のクラスタリングには、例えば音韻によるクラスタリングを行うことができる。

【0045】そして、入力音声素片 S_j 、およびトレーニング音声素片 T_{ij} のうち音韻が一致する音声素片のみをそれぞれ用いて、図2のステップS21、S22、S23、S24と同様の合成音声素片生成ステップS31、歪み評価ステップS32、代表音声素片生成ステップS33、音素環境クラスタ生成ステップS34の処理を順次行い、全ての初期音素環境クラスタについて同様の操作を繰り返すことにより、全ての代表音声素片およびそれに対応する音素環境クラスタの生成を行う。こうして生成された代表音声素片および音素環境クラスタは、図1の代表音声素片記憶部12および音素環境クラスタ記憶部13にそれぞれ記憶される。

【0046】ただし、各初期音素環境クラスタ当たりの

$$E_{c2} = \sum_{k=1}^N \min\{f(k, 1), f(k, 2), f(k, 3), \dots, f(k, N)\} \quad (3)$$

$$f(k, j) = \sum_{P_i \in C_k} e_{ij} \quad (4)$$

【0049】次に、代表音声素片生成ステップS44において、歪み e_{ij} に基づいて音素環境クラスタ C_k のそれぞれに対応する代表音声素片 D_k を入力音声素片 S_j より選択する。この代表音声素片 D_k は、入力音声素片 S_j から例えば次式(5)で表される歪み評価関数 E_{D2} を

$$E_{D2}(j) = \sum_{P_i \in C_k} e_{ij} \quad (5)$$

【0051】なお、この第3の処理手順による代表音声素片生成処理を変形し、第2の処理手順と同様に、何らかの先見的な知識に基づいて予め生成した初期音素環境クラスタ毎に代表音声素片の生成および音素環境クラスタの生成を行うことも可能である。

【0052】(第2の実施形態)次に、図5～図9を用いて本発明の第2の実施形態について説明する。図5は、第2の実施形態に係る音声合成方法を実現する音声合成装置の構成を示すブロック図である。図1と相対する部分に同一の参照符号を付して相違点を中心に説明すると、本実施形態では音声合成部15の後段に適応ポストフィルタ16が追加されている点が第1の実施形態と異なり、これに加えて代表音声素片生成部11における複数の合成音声素片の生成法も先の実施形態と異なっている。

【0053】すなわち、代表音声素片生成部11では第1の実施形態と同様に、トレーニング音声素片101にラベル付けされた音素環境102に含まれるピッチ周期

代表音声素片数が1であれば、初期音素環境クラスタが代表音声素片の音素環境クラスタとなるため、音素環境クラスタ生成ステップS34は不要となり、初期音素環境クラスタを音素環境クラスタ記憶部13に記憶すればよい。

【0047】次に、図4のフローチャートを参照して代表音声素片生成部11の第3の処理手順を説明する。この第3の処理手順による代表音声素片生成処理では、図2に示した第1の処理手順と同様に音声合成ステップS41および歪み評価ステップS42を順次経た後、次の音素環境クラスタ生成ステップS43において、音素環境 P_i および歪み e_{ij} に基づいて音素環境に関するクラスタ C_k ($k=1, 2, 3, \dots, N$)を生成する。音素環境クラスタ C_k は、例えば次式(3)(4)で表わされるクラスタリングの評価関数 E_{c2} を最小化するクラスタを探索することによって得られる。

【0048】

【数3】

(j)を最小化する音声素片を探索することによって得られる。

【0050】

【数4】

および継続時間長の情報に従って、入力音声素片103のピッチ周期および継続時間長を変更することで複数の合成音声素片を内部的に生成した後、これらの合成音声素片に対して適応ポストフィルタによるフィルタリングを施してスペクトル整形を行う。そして、この適応ポストフィルタによりスペクトル整形を行った後の各合成音声素片とトレーニング音声素片101との距離尺度に従って、代表音声素片104と音素環境クラスタ105が生成される。音素環境クラスタ105は、先の実施形態と同様にトレーニング音声素片101を音素環境に関するクラスタに分類して生成される。

【0054】なお、この代表音声素片生成部11において音素環境102に含まれるピッチ周期および継続時間長の情報に従って入力音声素片103のピッチ周期および継続時間長を変更して生成される複数の合成音声素片に対してフィルタリングを施してスペクトル整形を行う適応ポストフィルタは、音声合成部15の後段に配置される適応ポストフィルタ16と同様の構成でよい。

【0055】一方、音声合成部15では第1の実施形態と同様に代表音声素片選択情報107に従って代表音声素片記憶部12より選択的に読み出された代表音声素片108に対し、韻律情報111に従ってピッチ周期および音韻継続時間長を変更するとともに、素片の接続を行って合成音声信号113を生成するが、本実施形態ではこの合成音声信号113がさらに適応ポストフィルタ16に入力され、ここで音質向上のためのスペクトル整形が行われた後、最終的な合成音声信号114が取り出される。

【0056】図6に、適応ポストフィルタ16の一構成例を示す。この適応ポストフィルタ16は、ホルマント強調フィルタ21とピッチ強調フィルタ22を縦続配置して構成される。

【0057】ホルマント強調フィルタ21は、代表音声素片選択情報107に従って代表音声素片記憶部12から選択的に読み出された代表音声素片108をLPC分析して得られるLPC係数に基づいて決定されるフィルタ係数に従って、音声合成部15から入力される合成音声信号113をフィルタリングすることにより、スペクトルの山の部分を強調する処理を行う。一方、ピッチ強調フィルタ22は、韻律情報111に含まれるピッチ周期に基づいて決定されるパラメータに従って、ホルマント強調フィルタ21の出力をフィルタリングすることにより、音声信号のピッチを強調する処理を行う。なお、ホルマント強調フィルタ21とピッチ強調フィルタ22の配置順序は逆であってもよい。

【0058】このような適応ポストフィルタ16の適用によりスペクトルが整形され、「めりはり」のある明瞭な音声再生可能な合成音声信号114が得られる。適応ポストフィルタ16としては図6に示した構成のものに限られず、音声符号化や音声合成の分野で用いられる公知の技術に基づく種々の構成を採用することが可能である。

【0059】このように本実施形態では、規則合成系2において音声合成部15の後段に適応ポストフィルタ16が配置される点を考慮して、合成単位学習系1においても代表音声素片生成部11で音素環境102に含まれるピッチ周期および継続時間長の情報に従って入力音声素片103のピッチ周期および継続時間長を変更して生成される複数の合成音声素片に対し、同様に適応ポストフィルタによるフィルタリングを行っている。従って、適応ポストフィルタ16を通した後の最終的な合成音声信号114と同様のレベルで、自然音声に対する歪みが小さくなるような代表音声素片を代表音声素片生成部11において生成できるため、さらに自然音声に近い合成音声生成することが可能となる。

【0060】次に、図5における代表音声素片生成部11の処理手順について具体的に説明する。図7、図8および図9のフローチャートは、図5における代表音声素

片生成部11の第1、第2および第3の処理手順を示している。図7、図8および図9では、先に説明した図2、図3および図4に示した処理手順における音声合成ステップS21、S31およびS41の後に、ポストフィルタリングステップS25、S36およびS45が追加されている。

【0061】ポストフィルタリングステップS25、S36およびS45では、前述した適応ポストフィルタによるフィルタリングを行う。すなわち、音声合成ステップS21、S31およびS41で生成された合成音声素片 G_i に対し、入力音声素片 S_i をLPC分析して得られるLPC係数に基づいて決定されるフィルタ係数に従ってフィルタリングを行うことにより、スペクトルの山の部分を強調するホルマント強調を行う。また、このホルマント強調後の合成音声素片に対し、さらにトレーニング音声素片 T_i のピッチ周期に基づいて決定されるパラメータに従ってフィルタリングを行うことにより、ピッチ強調を行う。

【0062】このようにして、ポストフィルタリングステップS25、S36およびS45において、スペクトル整形を行う。このポストフィルタリングステップS25、S36およびS45は、前述したように規則合成系2において音声合成部15の後段に設けられる適応ポストフィルタ16により合成音声信号113のスペクトル整形を行って音質の向上を図るポストフィルタリングを行うことを前提に、合成単位の学習を可能とする処理であり、この処理を適応ポストフィルタ16による処理と組み合わせることによって、最終的に「めりはり」のある明瞭な合成音声信号114が生成される。

【0063】(第3の実施形態)次に、図10～図12を用いて本発明の第3の実施形態を説明する。図10は、第2の実施形態に係る音声合成装置における合成単位学習系の構成を示すブロック図である。

【0064】本実施形態における合成単位学習系30は、LPC分析・逆フィルタ部31、音源信号記憶部32、LPC係数記憶部33、音源信号生成部34、合成フィルタ35、歪み計算部36および最小歪探索部37からなる。この合成単位学習系30には、トレーニング音声素片101と、これにラベル付けされた音素環境102および入力音声素片103が入力される。入力音声素片103は、LPC分析・逆フィルタ部31に入力され、LPC分析が行われてLPC係数201と予測残差信号202が出力される。LPC係数201はLPC係数記憶部33に記憶され、また予測残差信号202は音源信号記憶部32に記憶される。

【0065】音源信号記憶部32に記憶された予測残差信号は、最小歪み探索部37からの指令に従って一つずつ読み出され、音源信号生成部34においてトレーニング音声素片101の音素環境102に含まれるピッチパターンおよび音韻継続時間長の情報に従って、そのピッ

10

20

30

40

50

チ周期および継続時間長が変更されることにより、音源信号が生成される。このようにして生成された音源信号は、最小歪探索部 3 7 からの指令に従って L P C 係数記憶部 3 3 から読み出された L P C 係数をフィルタ係数とする合成フィルタ 3 5 に入力され、合成音声素片が作成される。

【0066】次に、歪計算部 3 6 においてトレーニング音声素片 1 0 1 に対する合成音声素片の誤差つまり歪みが計算され、この歪みが最小歪探索部 3 7 において評価される。最小歪探索部 3 7 は、L P C 係数記憶部 3 3 と音源信号記憶部 3 2 にそれぞれ記憶されている L P C 係数と予測残差信号の全ての組み合わせを出力するように指令を出して、それらの組み合わせに対応して合成フィルタ 3 5 で合成音声素片を生成させる。そして、最小の歪みを与える L P C 係数と予測残差信号の組み合わせを見出し、それを記憶する。

【0067】次に、この合成単位学習系 3 0 の動作を図 1 1 のフローチャートを用いて説明する。まず、準備段階として連続発声された多数の音声データに音韻毎にラベリングを行い、C V, V C V, C V C などの合成単位に従って、トレーニング音声素片 T_i ($i = 1, 2, 3, \dots, N_T$) を切り出す。また、各トレーニング音声素片 T_i に対応する音素環境 P_i ($i = 1, 2, 3, \dots, N_T$) を抽出しておく。ただし、 N_T はトレーニング音声素片 T_i の個数を表す。音素環境は、少なくとも当該トレーニング音声素片の音韻とそのピッチパターンおよび継続時間長を含むものとし、その他に必要に応じて前後の音素などを含むものとする。

【0068】次に、トレーニング音声素片を作成したのと同様の方法により、多数の入力音声素片 S_i ($i = 1, 2, 3, \dots, N_S$) を作成する。ただし、 N_S は入力音声素片 S_i の個数を表す。ここで、入力音声素片 S_i とトレーニング音声素片 T_i の合成単位は一致させるものとする。例えば、ある C V 音節 “k a” の代表音声素片を作成する場合、多数の音声データから切り出された音節 “k a” から入力音声素片 S_i とトレーニング音

$$E_k(i, j) = D(T_k, G_k(i, j))$$

$$E_k(i, j) = \sum_{k=1}^{N_T} E_k(i, j)$$

【0075】ここで D は歪み関数であり、何らかのスペクトル距離を用いることができる。例えば、F F T などを用いてパワースペクトルを求めて、その間の距離を求める方法や、あるいは線形予測分析を行って L P C または L S P パラメータなどを求めて、パラメータ間の距離を評価する方法などがある。その他にも、短時間フーリエ変換やウェーブレット変換などの変換係数を用いて評価する方法が考えられる。また、各素片のパワーを正規

音声素片 T_i を設定する。なお、入力音声素片はトレーニング音声素片と同じものを使用してもよいし（すなわち、 $T_i = S_i$ ）、異なる音声素片を作成してもよい。いずれにしても、豊富な音韻環境を有する多数のトレーニング音声素片と入力音声素片が用意されていることが望ましい。

【0069】このような準備段階の後、L P C 分析ステップ S 5 1 で入力音声素片 S_i ($i = 1, 2, 3, \dots, N_S$) の L P C 分析を行い、L P C 係数 a_i ($i = 1, 2, 3, \dots, N_S$) を求めると共に、その係数に基づいた逆フィルタリングを行い、予測残差信号 e_i ($i = 1, 2, 3, \dots, N_S$) を求める。ただし、 a は p を L P C 分析の次数とすると、 p 個の要素を持つベクトルである。

【0070】次に、求められた予測残差信号を音源信号として、L P C 係数とともにステップ S 5 2 で保存する。次の L P C 係数・音源信号組み合わせステップ S 5 3 では、保存された L P C 係数と音源信号の組み合わせを一組 (a_i, e_j) 作成する。

【0071】この一組の組み合わせに対して、次の音声合成ステップ S 5 4 で P_k のピッチパターンおよび継続時間長に等しくなるように e_j のピッチおよび継続時間長を変更して音源信号を生成した後、L P C 係数 a_i を持つ合成フィルタでフィルタリング演算を行い、合成音声素片 $G_k(i, j)$ を生成する。

【0072】このように、全ての P_k ($k = 1, 2, 3, \dots, N_T$) に従って音声合成を行うことにより、 N_T 個の合成音声素片 $G_k(i, j)$ ($k = 1, 2, 3, \dots, N_T$) を生成する。

【0073】次の歪み評価ステップ S 5 5 では、合成音声素片 $G_k(i, j)$ とトレーニング音声素片 T_k との間の歪み $E_k(i, j)$ と P_k に関する歪みの総和 E を次式 (6) (7) により求める。

【0074】

【数 5】

(6)

(7)

化した上で、歪みの評価を行うことも考えられる。

【0076】ステップ S 5 3 ~ S 5 5 の処理を L P C 係数と音源信号の全ての組み合わせ (a_i, e_j) ($i, j = 1, 2, 3, \dots, N_S$) について行い、歪み評価ステップ S 5 5 で $E(i, j)$ の最小値を与える i, j の組を探索する。

【0077】次の代表音声素片生成ステップ S 5 7 では、 $E(i, j)$ の最小値を与える i, j の組、また

は、対応する (a_i, e_j) 、または、 (a_i, e_j) から生成される波形を代表音声素片として保存する。ただし、この代表音声素片生成ステップは、代表音声素片を各合成単位毎に一組生成する場合の処理であり、N組生成したい場合は、次のようにする。まず、 $N_s * N_s$

$$U = \{ (a_i, e_j)^n, m=1, 2, \dots, N \}$$

$$ED(U) = \sum_{k=1}^{N_T} \min(E_k(i, j)^n, E_k(i, j)^2, \dots, E_k(i, j)^n) \quad (9)$$

【0079】ただし、 $\min()$ は最小値を表す関数である。集合Uの組合せは、 $N_s * N_s$ C、通りあり、これらの集合Uの中から評価関数ED(U)を最小にするUを探索し、その要素 $(a_i, e_j)^k$ を代表音声素片とする。

【0080】次に、本実施形態における規則合成系について図12を用いて説明する。本実施形態における規則合成系40は、組み合わせ記憶部41、音源信号記憶部42、LPC係数記憶部43、音源信号生成部44および合成フィルタ45からなる。規則合成部40には、入力されたテキストの言語処理とそれに続く音韻処理の結果得られる韻律情報111と音韻記号列112が入力される。組み合わせ記憶部41、音源信号記憶部42およびLPC係数記憶部43には、図10の合成単位学習部30で求められたLPC係数と音源信号の組み合わせ情報 (i, j) と、音源信号 e_j 、LPC係数 a_i が予め記憶されている。

【0081】組み合わせ記憶部41は、音韻記号列112を入力し、これに対応する合成単位（例えばCV音節）を与えるLPC係数と音源信号の組合せ情報を出力する。音源信号記憶部42に記憶された音源信号は、組み合わせ記憶部41からの指令に従って読み出され、音源信号生成部44において入力された韻律情報111に含まれるピッチパターンおよび音韻継続時間長の情報に従って、そのピッチ周期および継続時間長が変更されると共に音源信号の接続が行われる。

【0082】こうして生成された音源信号は、組み合わせ記憶部41の指令に従ってLPC係数記憶部43から読み出された係数をフィルタ係数とする合成フィルタ45に入力され、フィルタ係数の補間とフィルタリング演算が行われることにより、合成音声信号113が作成される。

【0083】（第4の実施形態）次に、図13および図14を用いて本発明の第4の実施形態を説明する。図13は本実施形態における合成単位学習系の概略構成を示す図であり、第3の実施形態の図10に示した合成単位学習系30にクラスタリング部38を付加した構成とな

個の (a_i, e_j) の組の中からN組選択した集合を式(8)と置き、歪みの総和を表す評価関数を式(9)のように定義する。

【0078】

【数6】

(8)

っている。本実施形態では、クラスタリング部38において何らかの先見的な知識に基づいて予め音素環境のクラスタリングを行い、各クラスタに対して代表音声素片を生成する点が第3の実施形態と異なる。クラスタリングとしては、例えば当該素片のピッチによるクラスタリングが考えられる。この場合、トレーニング音声素片101をピッチに基づいてクラスタリングし、各クラスタのトレーニング音声素片に対して第3の実施形態で述べた代表音声素片の生成を行う。

【0084】図14は、本実施形態における規則合成系の概略構成を示す図であり、第3の実施形態の図12に示した規則合成系40にクラスタリング部48を付加した構成となっている。韻律情報111をトレーニング音声素片と同様にピッチクラスタリングし、合成単位学習系30で求められた各クラスタの代表音声素片に対応する音源信号及びLPC係数を用いて音声を作成する。

【0085】（第5の実施形態）次に、図15～図17を用いて本発明の第5の実施形態を説明する。図15は、本実施形態における合成単位学習系を示すブロック図であり、クラスタをトレーニング音声素片との歪み尺度に基づいて自動的に生成する場合の構成例を示している。本実施形態は、図10に示した合成単位学習系30に音素環境クラスタ生成部51とクラスタ記憶部52が追加された構成となっている。

【0086】本実施形態における合成単位学習系の第1の処理手順を図16に示すフローチャートを用いて説明する。この処理手順は第3の実施形態の処理手順を示した図11に新たに音素環境クラスタ生成ステップS58が追加されている。このステップS58では、音素環境P、と歪み $E_k(i, j)$ および代表音声素片D、より、音素環境に関するクラスタC、 $(m=1, 2, 3, \dots, N)$ を生成する。音素環境クラスタC、は、例えば次式(10)で表されるクラスタリングの評価関数 E_c を最小化するクラスタを探索することによって得られる。

【0087】

【数7】

$$E_{c_m} = \sum_{n=1}^N \sum_{P_k \in C_m} E_k(i, j)$$

(10)

【0088】図17は、図15の合成単位学習系の第2の処理手順を示すフローチャートである。この処理では、初期音素環境クラスタ生成ステップS50で何らかの先見的な知識に基づいて予め音素環境のクラスタリングを行い、初期音素環境クラスタを生成する。クラスタリングとしては、例えば当該素片の音韻によるクラスタリングが考えられる。この場合、当該素片の音韻が一致する音声素片およびトレーニング音声素片だけを用いて第3の実施形態で述べた代表音声素片の生成および音素環境クラスタの生成を行い、全ての初期音素環境クラスタについて同様の操作を繰り返すことによって、全ての代表音声素片および対応する音素環境クラスタの生成を行う。

【0089】ただし、各初期クラスタ当たりの代表音声素片数が1であれば、初期音素環境クラスタが代表音声素片の音素環境クラスタとなるため、音素環境クラスタ生成ステップS58は不要となり、初期音素環境クラスタを図15のクラスタ記憶部52に記憶すればよい。

【0090】本実施形態における規則合成系は、図14に示した第4の実施形態における規則合計系40と同様に構成される。この場合、クラスタリング部48は図15のクラスタ記憶部52に蓄積された情報に基づいて処理を行う。

【0091】（第6の実施形態）図18に、本発明の第6の実施形態における合成単位学習系の構成を示す。本実施形態における合成単位学習系は、図10に示した合成単位学習系30にバッファ61、62および量子化テーブル作成部63、64が追加された構成となっている。

【0092】本実施形態において、入力音声素片103はLPC分析・逆フィルタ部31に入力され、ここでLPC分析により生成されたLPC係数201と予測残差信号202が一旦バッファ61、62にそれぞれ蓄えられた後、量子化テーブル作成部63、64でそれぞれ量子化され、量子化されたLPC係数と予測残差信号がLPC係数記憶部33および音源信号記憶部34にそれぞれ記憶される。

【0093】図19は、図18の合成単位学習系の処理手順を示すフローチャートであり、図11のフローチャートに示した処理手順と異なるところは、LPC分析ステップS51の後に量子化ステップS60が追加されたことである。この量子化ステップS60では、LPC分析ステップS51で求められたLPC係数 a_i ($i = 1, 2, 3, \dots, N_s$) と予測残差信号 e_i ($i = 1, 2, 3, \dots, N_s$) を一旦バッファに蓄積した後、LBGアルゴリズムなどの公知の技術を用いて量子化テーブルを作成し、LPC係数と予測残差信号を量子化する。

このとき、量子化テーブルのサイズ、すなわち量子化の代表ベクトルの数は N_s 未満とする。そして、量子化されたLPC係数と予測残差信号が次のステップS52で保存される。その後の処理は、図11の場合と同一である。

【0094】（第7の実施形態）図20は、本発明の第7の実施形態における合成単位学習系を示すブロック図であり、クラスタをトレーニング音声素片との歪み尺度に基づいて自動的に生成する場合の構成例を示している。クラスタの生成は、第5の実施形態と同様に行うことができる。すなわち、本実施形態における合成単位学習系は、図15に示した第5の実施形態と図18に示した第6の実施形態とを組み合わせた構成となっている。

【0095】（第8の実施形態）図21は、本発明の第8の実施形態における合成単位学習系であり、LPC分析部31aと逆フィルタ31bを分離して、バッファ61および量子化テーブル作成部63を経て量子化されたLPC係数を用いて逆フィルタリングを行って予測残差信号を計算する場合の構成例を示している。このようにすることにより、LPC係数の量子化歪みによる合成音声の音質劣化を低減する代表音声素片を生成することが可能になる。

【0096】（第9の実施形態）図22は、本発明の第9の実施形態における合成単位学習系であり、第8の実施形態と同様に、量子化されたLPC係数を用いて逆フィルタリングし、予測残差信号を計算する場合の他の構成例を示している。ただし、本実施形態では逆フィルタ31bで逆フィルタリングされた予測残差信号がバッファ62および量子化テーブル64を経て量子化された後、音源信号記憶部32に入力される点が第8の実施形態と異なっている。

【0097】第6～第9の実施形態において、量子化テーブル作成部63、64で作成される量子化テーブルのサイズ、すなわち量子化の代表スペクトルの数は、クラスタ数または合成単位の総数（例えば、CV、VC音節の総数）より少なく選ぶことができる。このようにLPC係数と予測残差信号を量子化することによって、代表音声素片として記憶されるLPC係数と音源信号の数を減少させることができるため、合成単位の学習に要する計算時間を短縮することができると共に、規則合成系で用いるメモリ量を低減することができる。

【0098】しかも、LPC係数と音源信号の組み合わせ (a_i, e_i) で音声合成を行うので、LPC係数と音源信号数のどちらかの代表音声素片数がクラスタ数や合成単位の総数（例えば、CV、VC音節の総数）より少ない場合でも、良好な合成音声を得ることができる。

【0099】また、第6～第9の実施形態において、ト

レーニング音声素片と合成音声素片との歪み尺度として合成素片間の接続歪みを考慮することにより、より滑らかな合成音を得ることもできる。

【0100】さらに、合成単位の学習および規則合成において、第2の実施形態で説明したと同様の適応ポストフィルタを合成フィルタと合わせて用いることもでき、これにより合成音声のスペクトルが整形され、「めりはり」のある明瞭な合成音を得ることができる。

【0101】

【発明の効果】以上説明したように、本発明の音声合成方法によれば、入力音声素片に対してピッチおよび継続時間長の少なくとも一方の変更を行って生成される合成音声のレベルで自然音声に対する歪みを評価し、それに基づいて入力音声素片から選択した音声素片を代表音声素片とするため、音声合成装置の特性をも考慮した代表音声素片の生成が可能であり、この代表音声素片を接続して音声合成を行うことによって、自然音声に近い高品質の合成音声を生成することができる。

【0102】また、本発明ではさらに代表音声素片の接続によって合成される音声に対してスペクトル整形を行うと共に、合成音声素片に対しても同様のスペクトル整形を行うことにより、スペクトル整形後の最終的な合成音声信号のレベルで、自然音声に対する歪が小さくなるような代表音声素片を生成できるため、「めりはり」のあるより明瞭な合成音声を生成することができる。

【0103】また、各代表音声素片を音素環境に基づく素片選択規則に従って選択して接続することにより、合成音声は滑らかで自然性の高いものとなる。さらに、代表音声素片として音源信号（例えば予測残差信号）音源信号を入力として合成音声信号を生成する合成フィルタの係数（例えばLPC係数）の組の情報を記憶する場合、これらを量子化することによって、代表音声素片として記憶する音源信号と合成フィルタの係数の数を減少させることができるため、合成単位の学習に要する計算時間を短縮することができると共に、規則合成系で用いるメモリ量を低減することができる。

【0104】しかも、代表音声素片の情報として記憶する音源信号および合成フィルタの係数のうちの少なくとも一方の数を音声合成単位の総数（例えば、CV、VC音節の総数）や音素環境クラスタ数より少ない場合でも、良好な合成音声を得ることができる。

【図面の簡単な説明】

【図1】本発明の第1の実施形態に係る音声合成装置の構成を示すブロック図

【図2】図1中の代表音声素片生成部での第1の処理手順を示すフローチャート

【図3】図1中の代表音声素片生成部での第2の処理手順を示すフローチャート

【図4】図1中の代表音声素片生成部での第3の処理手順を示すフローチャート

【図5】本発明の第2の実施形態に係る音声合成装置の構成を示すブロック図

【図6】図5中の適応ポストフィルタの構成例を示すブロック図

【図7】図5中の代表音声素片生成部での第1の処理手順を示すフローチャート

【図8】図5中の代表音声素片生成部での第2の処理手順を示すフローチャート

【図9】図5中の代表音声素片生成部での第3の処理手順を示すフローチャート

【図10】本発明の第3の実施形態に係る音声合成装置における合成単位学習系の構成を示すブロック図

【図11】図10の合成単位学習系の処理手順を示すフローチャート

【図12】本発明の第3の実施形態に係る音声合成装置における規則合成系の構成を示すブロック図

【図13】本発明の第4の実施形態に係る音声合成装置における合成単位学習系の構成を示すブロック図

【図14】本発明の第4の実施形態に係る音声合成装置における規則合成系の構成を示すブロック図

【図15】本発明の第5の実施形態に係る音声合成装置における合成単位学習系の構成を示すブロック図

【図16】図15の合成単位学習系の第1の処理手順を示すフローチャート

【図17】図15の合成単位学習系の第2の処理手順を示すフローチャート

【図18】本発明の第6の実施形態に係る音声合成装置における合成単位学習系の構成を示すブロック図

【図19】図18の合成単位学習系の処理手順を示すフローチャート

【図20】本発明の第7の実施形態に係る音声合成装置における合成単位学習系の構成を示すブロック図

【図21】本発明の第8の実施形態に係る音声合成装置における合成単位学習系の構成を示すブロック図

【図22】本発明の第9の実施形態に係る音声合成装置における合成単位学習系の構成を示すブロック図

【符号の説明】

1…合成単位学習系

2…規則合成系

11…代表音声素片生成部

12…音素環境クラスタ記憶部

13…代表音声素片記憶部

14…素片選択部

15…音声合成部

16…適応ポストフィルタ

21…ホルマント強調フィルタ

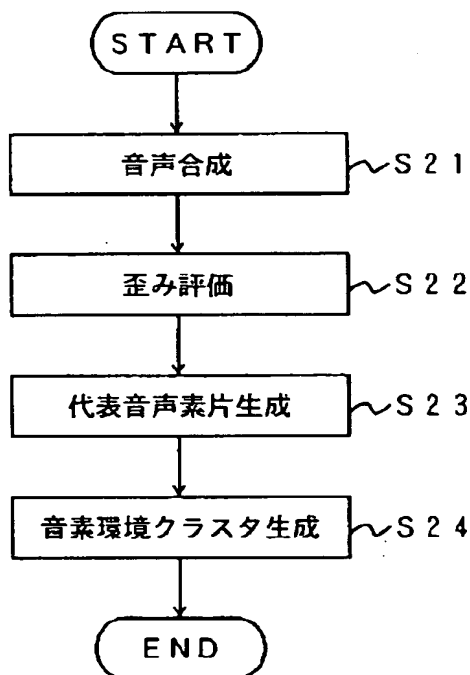
22…ピッチ強調フィルタ

101…トレーニング音声素片（第1の音声素片）

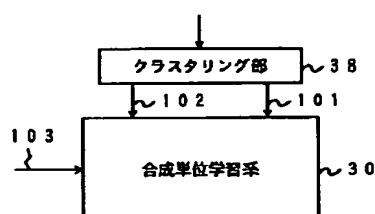
102…トレーニング音声素片にラベル付けされた音素環境

103…入力音声素片 (第2の音声素片)
 104…代表音声素片
 105…音素環境クラスタ
 106…音素環境クラスタ
 107…代表音声素片選択情報
 108…代表音声素片
 111…韻律情報
 112…音韻記号列
 113…合成音声信号
 114…合成音声信号
 30…合成単位学習系
 31…LPC分析・逆フィルタ
 31a…LPC分析部
 31b…逆フィルタ
 32…音源信号記憶部
 33…LPC係数記憶部

【図2】

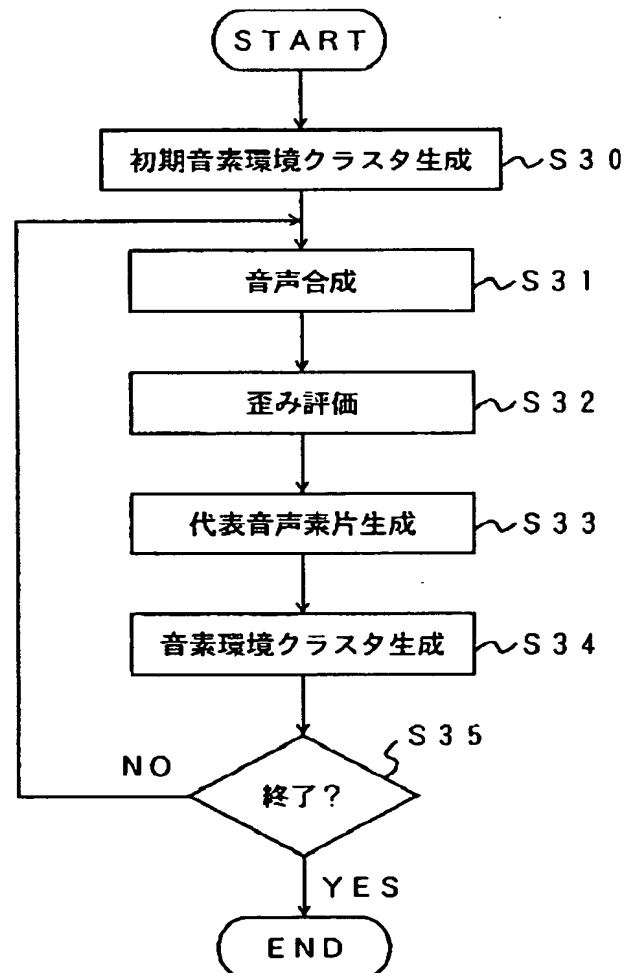


【図13】

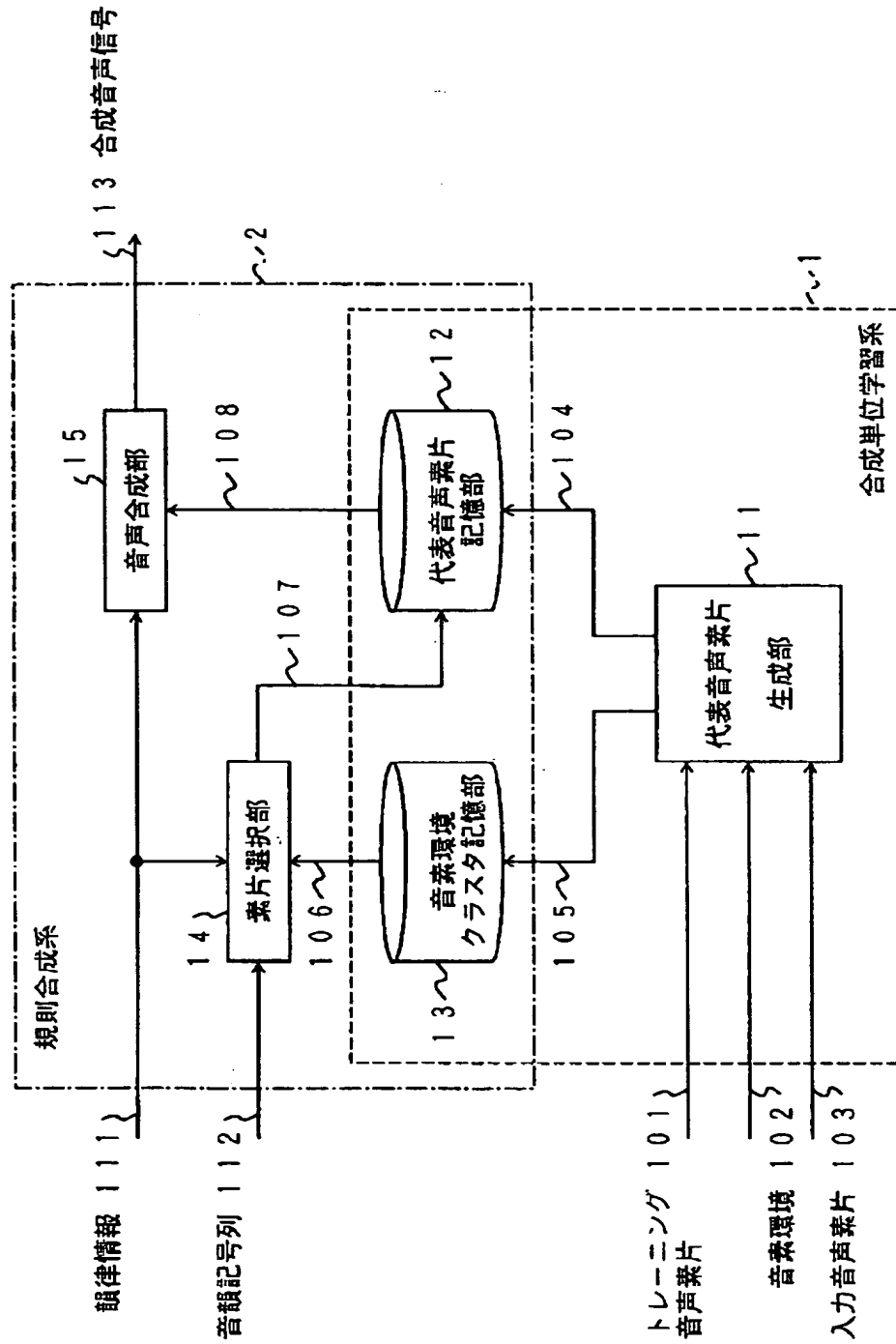


34…音源信号生成部
 35…合成フィルタ
 36…歪計算部
 37…最小歪探索部
 38…クラスタリング部
 40…規則合成系
 41…組み合わせ記憶部
 42…音源信号記憶部
 43…LPC係数記憶部
 10 44…音源信号生成部
 45…合成フィルタ
 48…クラスタリング部
 51…音素環境クラスタ生成部
 52…クラスタ記憶部
 61, 62…バッファ
 63, 64…量子化テーブル作成部

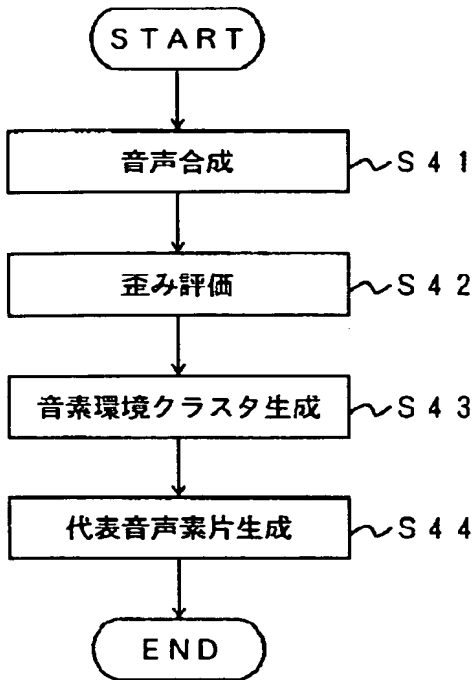
【図3】



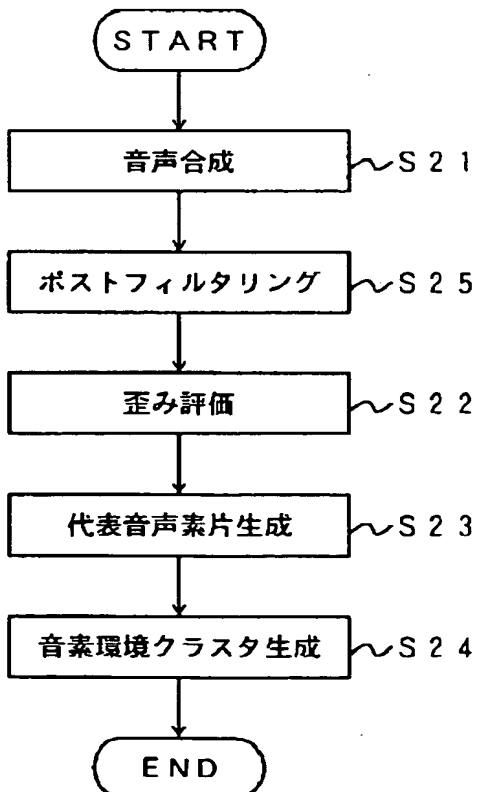
【図 1】



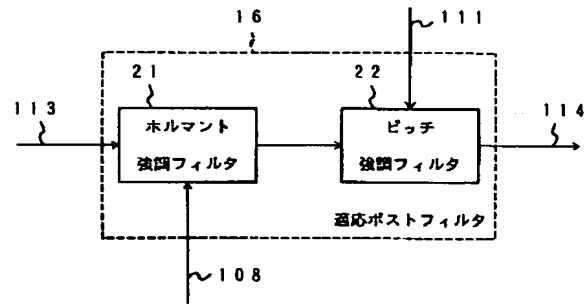
【図 4】



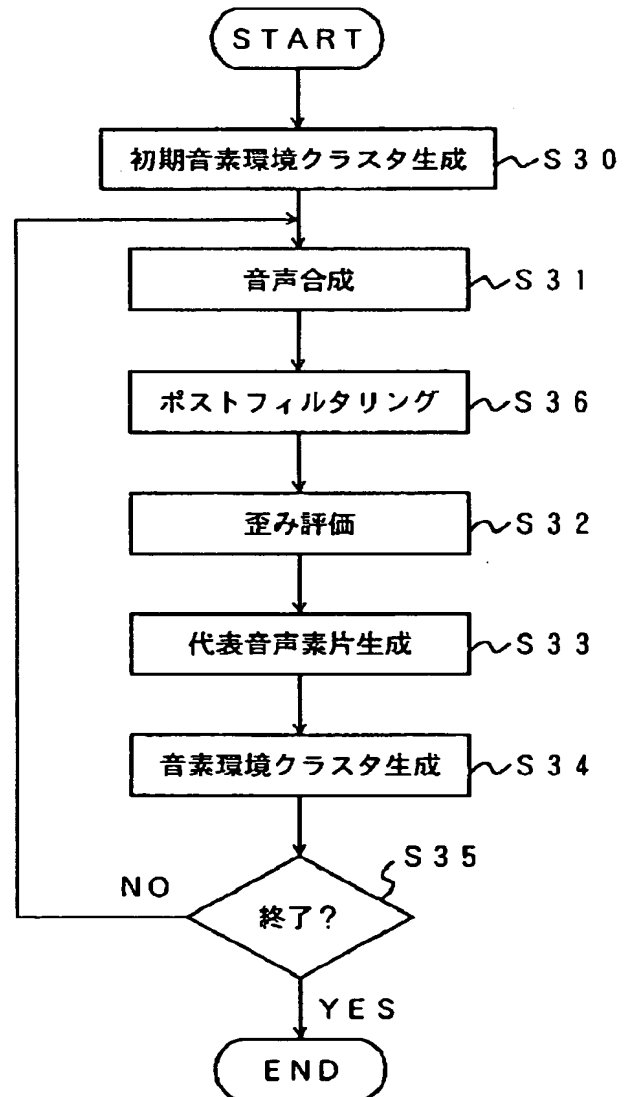
【図 7】



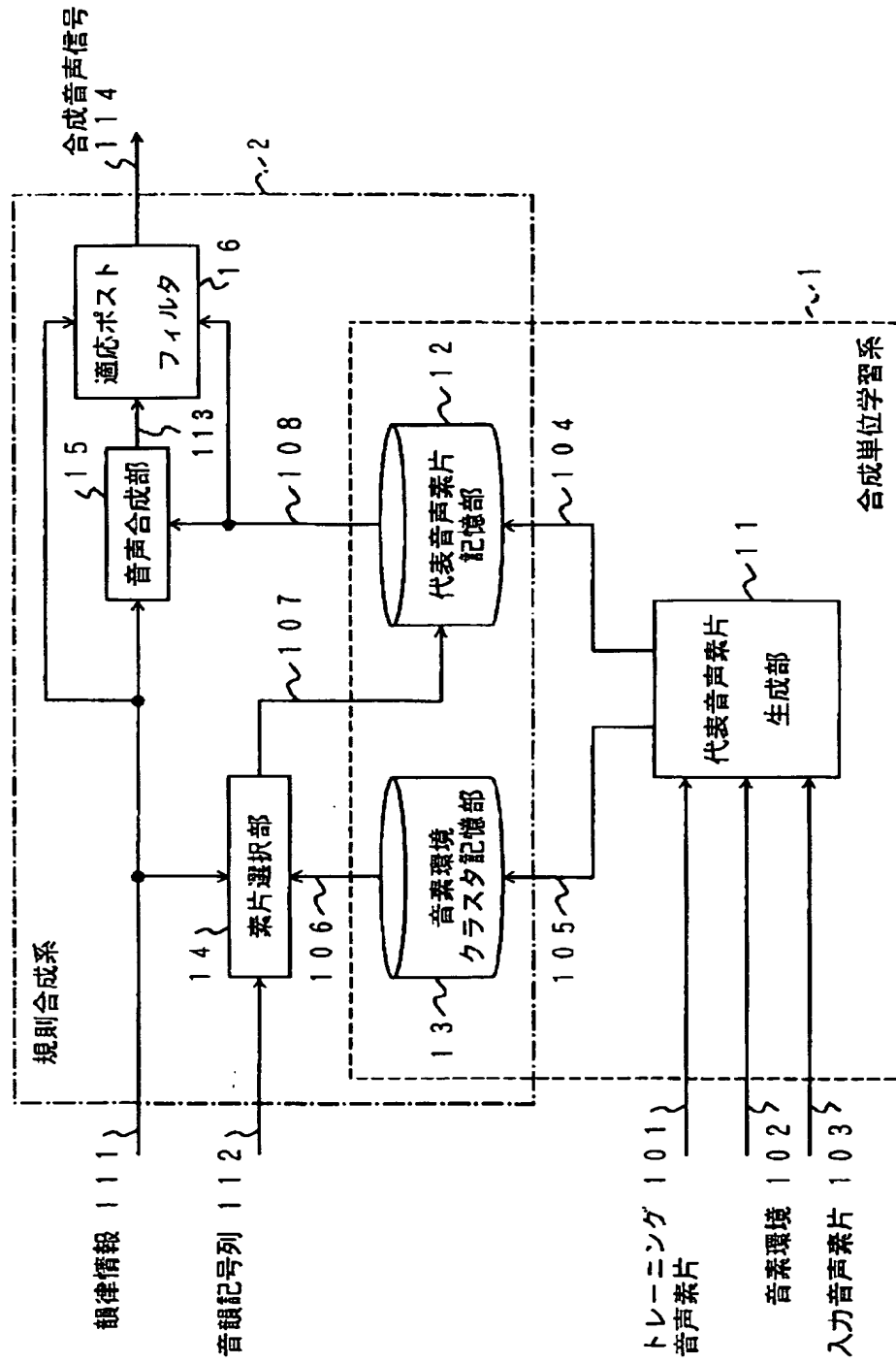
【図 6】



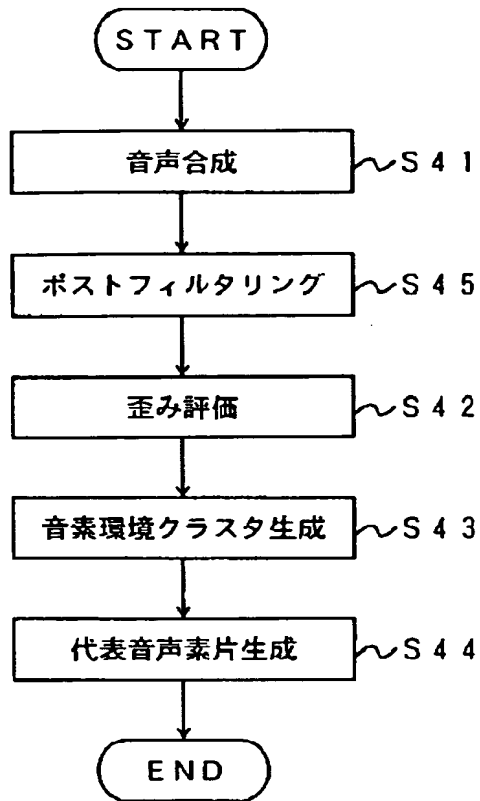
【図 8】



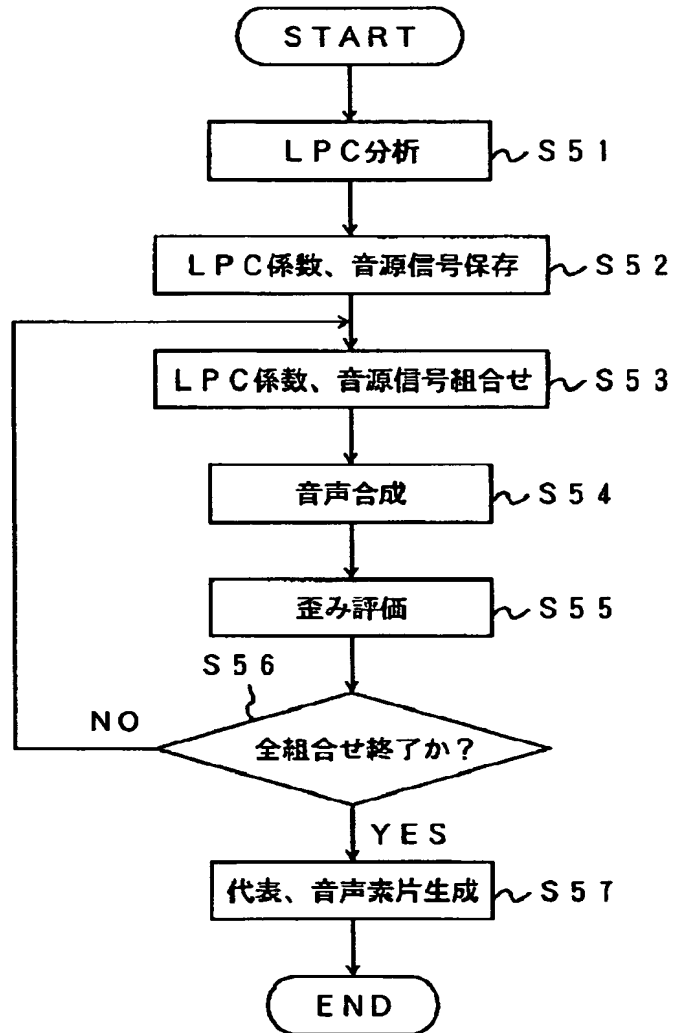
【図 5】



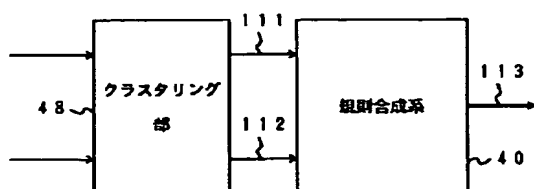
【図 9】



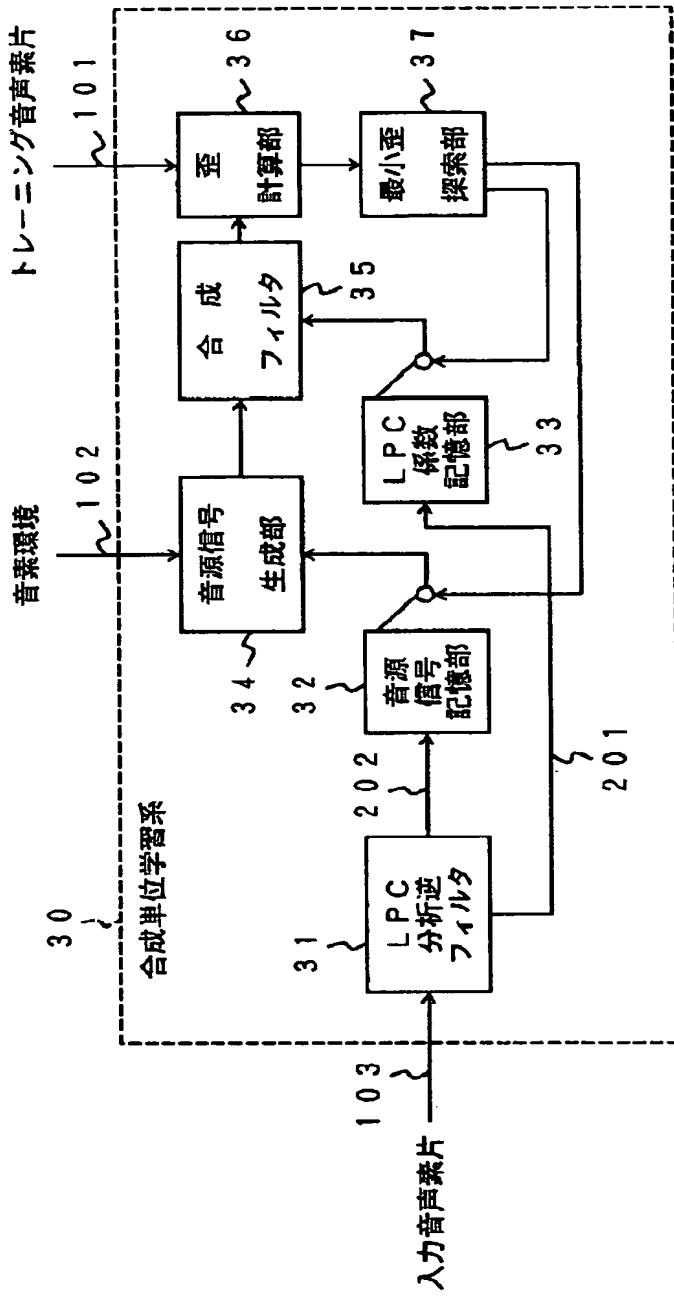
【図 11】



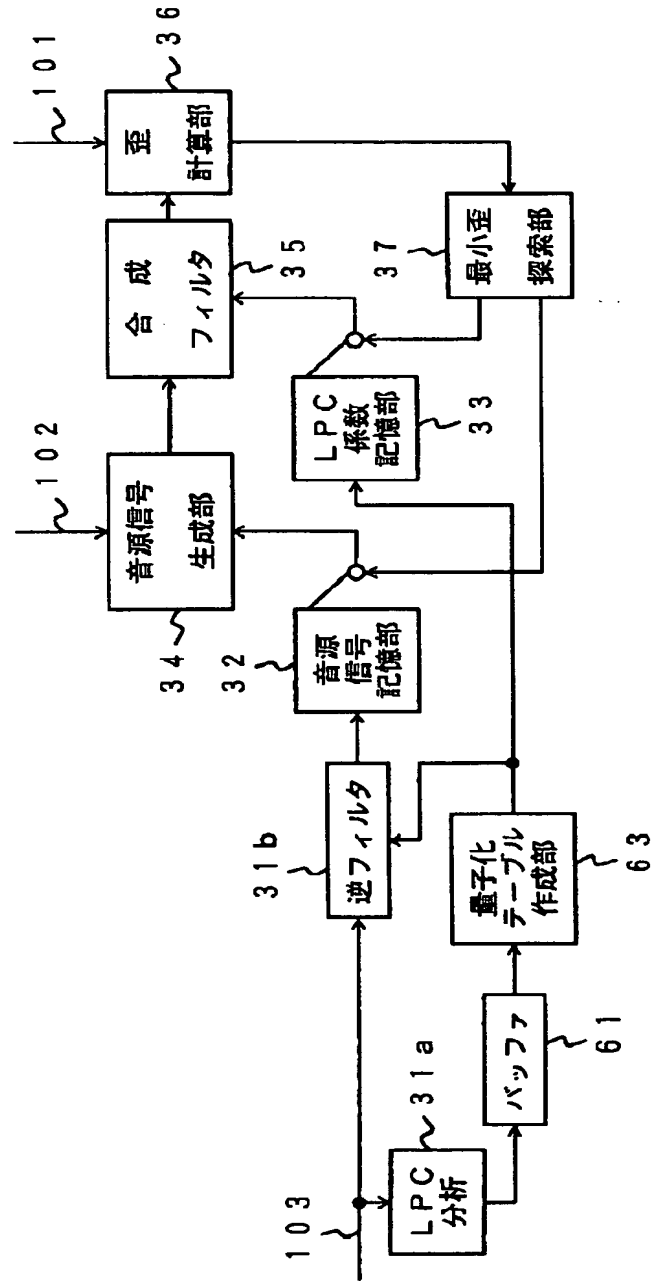
【図 14】



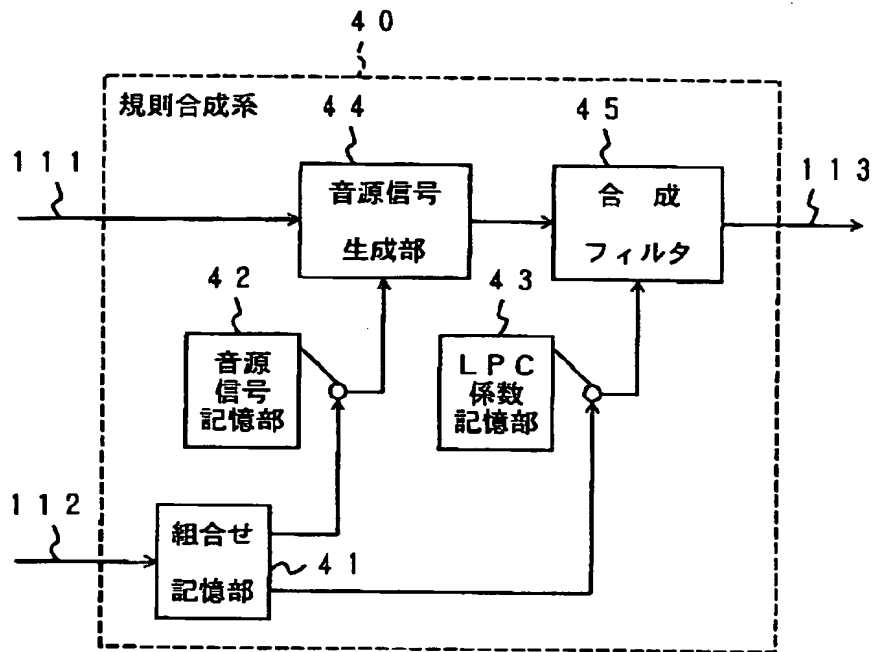
【図 10】



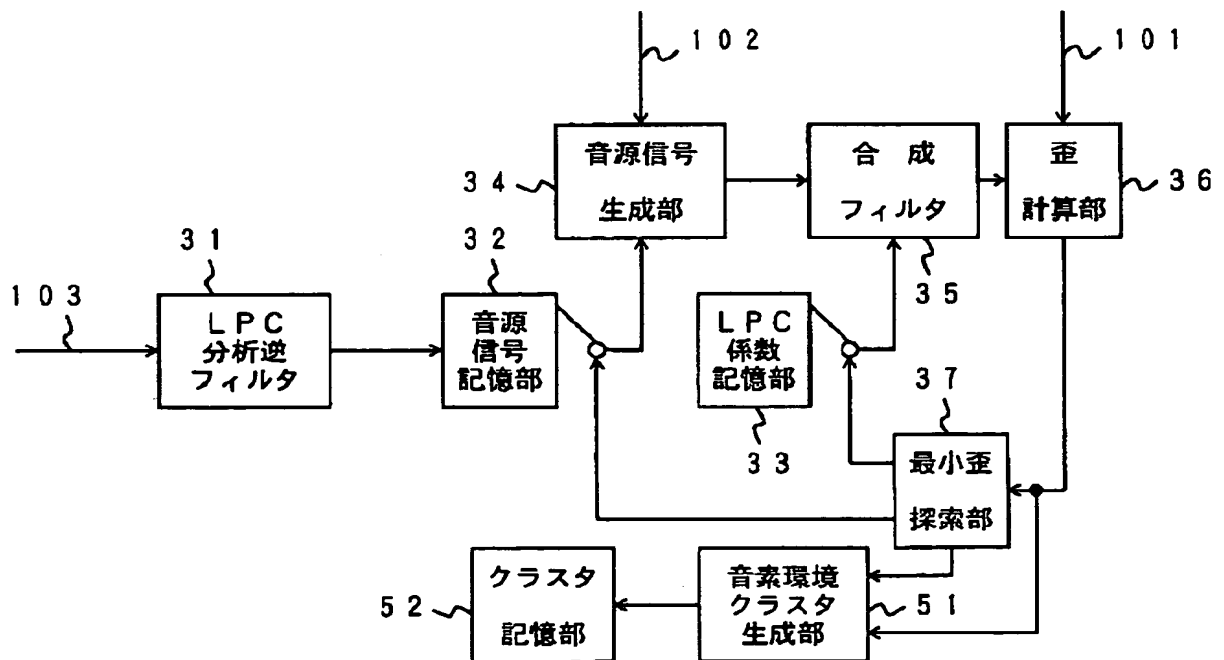
【図 21】



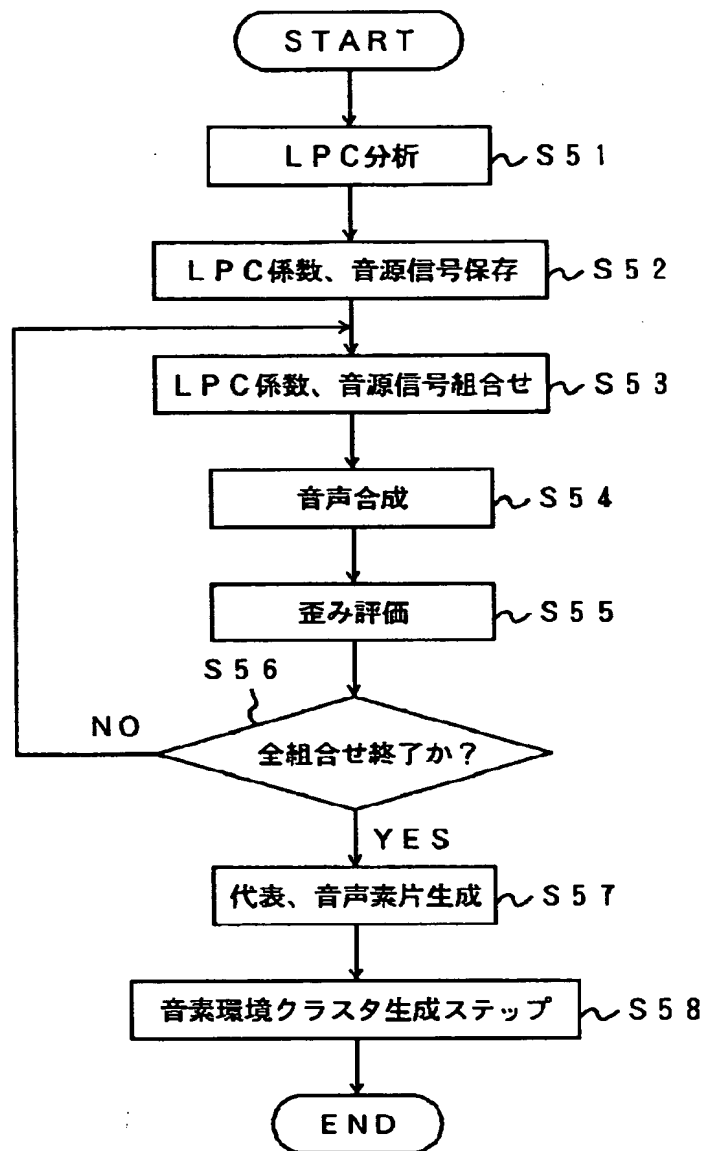
【図 1 2】



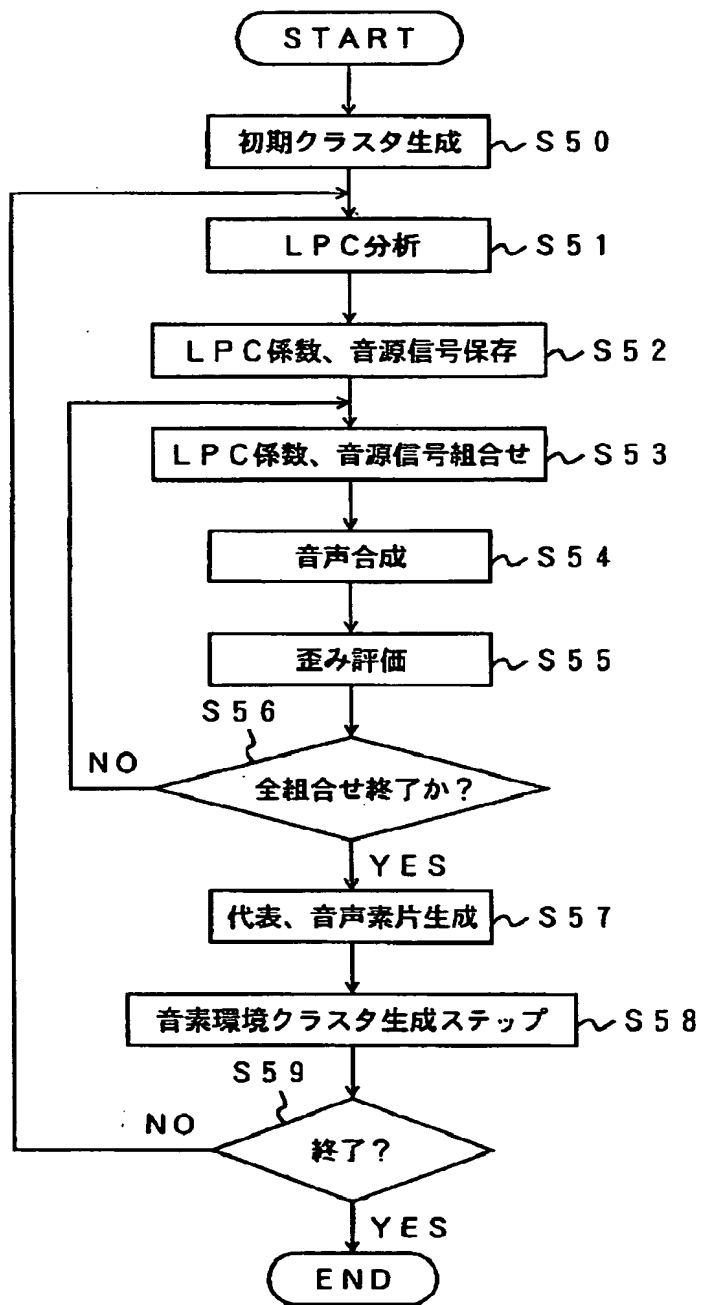
【図 1 5】



【図 16】



【図 17】



【図 19】

